Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits

ROBERT H. SHUMWAY,*,† RAHMAN S. AZARI,† AND MASOUD KAYHANIAN

Department of Statistics and Department of Civil and Environmental Engineering, University of California, Davis, California 95616

We review statistical methodology for estimating mean concentrations of potentially toxic pollutants in water, for small samples that are not normally distributed and often contain substantial numbers of nondetects, i.e. samples that are only known to be below some set of fixed thresholds. Maximum likelihood estimation (MLE) and regression on order statistics (ROS) are two main approaches that dominate the literature, with transformation bias under non-normality that increases with the severity of censoring being the main problem. We consider exact maximum likelihood estimators in conjunction with the Box-Cox transformation and propose the Quenouille-Tukey Jackknife as a method for bias reduction and variance estimation. Exact maximum likelihood estimators resulting from the expectation-maximization (EM) algorithm are exhibited in a simple heuristic form that also provides estimated values for the nondetects as subsidiary outputs. We show in simulations that the two main approaches perform well for the log-normal and gamma distributions as long as the jackknife is employed to reduce bias. Bias corrections to MLE used in the literature are shown to correct in the wrong direction under severe censoring. The jackknife is also used for estimating the variance of the both the MLE and ROS estimators. Robustness is improved by searching a class of power transformations (Box-Cox) for the best approximating normal distribution. We conclude that both the exact MLE and ROS procedures can be useful under varying experimental conditions. Limited simulations indicate that the ROS procedure is unbiased and has a smaller variance than the MLE under the log-normal distribution and is robust. The MLE performed better in simulations involving the gamma as the underlying distribution. We also compare the estimators for the mean and variance that one obtains from typical sets of water quality data, analyzing for copper, alumnium, arsenic, chromium, nickel, and lead.

Introduction

The problem of estimating parameters of the normal distribution under censoring or truncation has a long history, dating back at least to the papers of Hald (*1*) who derived the maximum likelihood estimators (MLEs) and their large-sample variance covariance matrix and Aitchison and Brown (*2*) who did the same for the log-normal distribution. The equations in their primitive versions could not be solved in closed form, and so application of iterative methods such as Newton−Raphson (see ref *3*), while theoretically feasible, were computationally intensive and often unstable for starting values chosen away from the maximum. This realization spawned a number of papers that advocated either the use of various computational tables such as in ref *4* or approximations to the exact likelihood equations such as in ref *5*. Such approximations persist to the present (see, for example, refs *6* and *7*) and even appear in software packages such as ref *8* when it is well-known that exact MLEs are available by simple and numerically stable iterations using the EM algorithm as in refs *9* and *10*.

A common concern in the literature has been for bias that is inherent in many maximum likelihood estimators and which is exacerbated by transformations designed to improve the accuracy of the normal likelihood with censored data. For non-normal distributions, it is common in water quality work to assume that the log-normal distribution applies so that the logarithms of the raw data can be used in a Gaussian likelihood. The means and variances of the transformed variables are related nonlinearly to the original means and variances, and the process of transforming back gives estimators that often are quite severely biased. Bias corrections for the conventional untransformed MLEs have been derived in ref *11* under Type II censoring, where observations are discontinued after a specified number of failures. The usual assumption in water quality problems is that nondetects are all below some threshold, a situation commonly referred to as Type I censoring. Schneider and Weissfield (*12*) continue this bias study using a least-squares fit to Saw's table to obtain a correction depending strictly on the percentage of censored observations, i.e., the correction depends on both the number of observations and the number of nondetects. Transformation bias introduced by the log-normal assumption was discussed by El-Shaarawi (*13*), who computed the expectation of the estimated untransformed means using the large-sample distributional properties of the maximum likelihood estimators in the transfomed space. The large sample properties substituted were for the MLEs assuming that there has been no censoring. Besides this use of an inappropriate asymptotic variance covariance matrix, there is the problem that the bias correction depends on the MLEs so that the distribution of the corrected estimator has a further nonlinearity. Simulation studies of the Saw bias correction in refs *14* and *6* showed that it failed to correct properly for samples with greater than 40% censoring. In this paper, we will see that the Saw bias correction actually adjusts a small bias in the wrong direction. We obtain some limited success, however, using the Quenouille-Tukey Jackknife (see ref *15*) as a method of correcting for bias and for estimating the variances of the estimated means.

Most methodology recognizes that it is unlikely that the underlying water quality data will be either normal or log-normally distributed and entertains various procedures for protecting against distributional departures. Two comprehensive simulation papers by Gilliom and Helsel (*16*) and Helsel and Gilliom (*17*) used regression on the normal scores of order statistics of the log transformed data. This was originally proposed by Gupta (*18*) for the untransformed case, who used regression weighted by the inverse of the covariance of the order statistics. Gilliom and Helsel evaluate eight different methods including maximum likelihood and conclude that the method of regression on order statistics (ROS) is best in terms of mean square error. They also assert superior

* Corresponding author phone: (530)752-6475; fax: (530)756-3404; e-mail: shumway@wald.ucdavis.edu.

† Department of Statistics.

‡ Department of Civil and Environmental Engineering.

10.1021/es0111129 CCC: $22.00 Published on Web 06/19/2002

coverage properties for confidence intervals computed by ROS over MLE intervals, but their study is flawed by using variance estimates computed from the simulations and not from theoretical properties of ROS or MLE. There should be an estimator for the variance of the ROS or MLE estimator that can be computed directly from the sample. We resolve this difficulty in the current paper by using the large-sample Cramér-Rao lower bound for the MLE and jackknife estimators for both the MLE and ROS methods.

The issue of robustness to underlying distribution can also be approached by estimating a transformation that leads to the best approximate Gaussian likelihood, as in ref *10*. Maximum likelihood estimators in the transformed space can be transformed back to obtain maximum likelihood estimators for the mean and variance in the original untransformed scale. A class of power transformations (see ref *19*) of the form

$$y_{\mathrm{i}}(\lambda) = \begin{cases} (x_i^{\lambda} - 1)/\lambda, & \lambda \neq 0 \\ \ln x_i, & \lambda = 0 \end{cases} \qquad (1)$$

for a sample x_i, $i = 1, 2, ..., n$ has potential, where the mean and variance of x_i, say μ_x,σ_x^2, expressed in terms of the moments of y_i, say μ_y,σ_y^2, are fairly easy to compute for the special cases $\lambda = 0, .50, 1$, corresponding to the logarithmic, square root, and no transformation, respectively, using eq 10. For limited censoring, ref *10* shows that searching the transformed likelihood produces an effective transformation in many cases and evaluates the effects on the coverage of the confidence intervals obtained through such a procedure. In particular, it is important not to apply a transformation such as the logarithmic form when one is not needed.

A number of the simulation studies (for example, refs *16*, *17*, and *6*) have investigated the effect of filling in various values for the unobserved elements of the sample that fell below the detection limits. Suggestions considered are replacing the values by zero, the detection limit, or one-half the detection limit. Such arbitrary procedures seem less defensible than the use of values following from theoretically defensible arguments. For normal variables, the EM algorithm for exact MLE (see ref *10*) uses the conditional mean, $E(y_i|y_i \leq T_i)$, and conditional variance of the normal distribution at each stage of the iterative procedure (see also ref *21*) as fill-in contributions for the components of the sample mean and variance. Another sensible fill-in procedure, used for the ROS estimators, is to fit a straight line to the normal scores of the order statistics for the observed values and then to fill in values extrapolated from the straight line for the observations below the detection limit. We concentrate on these latter two methods in this investigation.

In the second section, we summarize the two recommended methodologies, namely, exact maximum likelihood estimation (MLE) and regression on order statistics (ROS), with additional contributions to the bias reduction and variance estimation problem using the Quenouille-Tukey Jackknife. The jackknife estimator has also been suggested previously by Singh et al. (*20*). We also illustrate the two recommended procedures on a set of contrived data, generated from the log-normal distributions. The third section shows simulations involving the various methods for samples of sizes $n = 20, 50$ with high levels of censoring (50%, 80%). In section 4, we illustrate results on a group of real metal concentrations in water samples.

Statistical Methodology

Suppose that we observe a sample that could conceptually contain values $x_1, x_2, ..., x_n$, with common mean μ_x and variance σ_x^2. In the Type I censored case, a subset of $n_0 < n$ values are only known to be below some threshold, i.e., $x_i \leq T_i$. The remaining n_1 values ($n_0 + n_1 = n$) are observed as x_i, $x_i > T_i$. We assume further that the population x values may be distinctly non-normal, as, for example, in the case of water quality data, which may be log-normally distributed or may have a gamma distribution. We assume that there is a transformation of the form (1) which produces an approximately normally distributed set of values y_i with transformed thresholds, U_i generated by applying (1) to the thresholds implied for the raw data. The methods given below for the MLE and ROS assume the above conditions.

Regression on Order Statistics (ROS). A number of potentially *robust methods* are available using the normal scores for the order statistics. We take the one here recommended in ref *8*. Suppose that a transformation (logarithmic or otherwise) has yielded n_0 observations, y_i, $i = 1, 2, ..., n_0$, each below a common transformed detection limit U and n_1 observations y_i, $i = n_0 + 1, ..., n_0 + n_1$ that are observed and greater than U. If the observations are independently normally distributed and have common mean μ_y and variance σ_y^2, the mean and variance will satisfy the equation

$$y_i = \mu_y + \sigma_y \Phi^{-1}(P_{\mathrm{i}})$$

where $P_{\mathrm{i}} = Prob\{Y_i \leq y_i\}$ and $\Phi^{-1}(\cdot)$ denotes the inverse of the cumulative normal distribution function 13, defined in the Appendix. This suggests that the intercept and slope from a regression on the normal scores would yield the mean and variance of the transformed observations. The regression is performed on the inverse transformed adjusted order statistics. It should be noted here that if the procedure is truly robust to departures from normality, the original untransformed observations x_{i} and detection limits T_i could be used. Our simulations, presented later in the third section, suggest that regressing on the order statistics is robust for log-normal populations but is distinctly nonrobust if the underlying data follow a gamma distribution.

The commonly accepted procedure is to replace the probabilities by the adjusted ranks (*22*), so that the regression equation becomes

$$y_i = \mu_y + \sigma_y \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right) + \epsilon_{\mathrm{i}} \qquad (2)$$

where $i = n_0 + 1, n_0 + 2, ..., n_0 + n_1$, with the estimators for μ_y and σ_y estimated by least squares. This implies that the residual errors ϵ_i are assumed to have equal variance and are uncorrelated. Gupta (*18*) gives the covariance matrix of the order statistics for relatively small samples and then uses weighted least squares. In the present case, Helsel and Gilliom (*17*) recommend using ordinary least squares as an easier computational alternative. Their procedure uses the predicted values from the regression model (2) for $i = 1, 2, ..., n_0$ for the censored values. Then, transform back to the original observations $\hat{x}_{\mathrm{i}}$ and compute the usual mean and variance from the resulting sample $\hat{x}_1, \hat{x}_2 ..., \hat{x}_{n0}, x_{n0+1}, ..., x_{n0+n1}$ for $\hat{\mu}_x$ and $\hat{\sigma}_x^2$. Note that the procedure produces estimators for the censored values based on extrapolation from a normal model for the transformed values and then back-transforming to the original raw observations.

The literature is not clear on what is used for the variance of $\hat{\mu}_x$ and how this can be incorporated into a 95% confidence interval. One possibility is to use the sample variance of the extrapolated sample and make the assumptions that are used in the uncensored case with an adjustment for the degrees of freedom. We have not followed that procedure here but have utilized the jackknife estimator described in the next section.

Bias, Variance, and the Jackknife. We may apply the jackknife for estimating the variance of the parameter estimators produced by the ROS procedure since no theoretical arguments have been given in the literature for preferring another method. Suppose that we have an estimator, generically denoted by $\hat{\theta}$, for some parameter θ. The jackknife estimator computes a collection of n pseudoestimators, by deleting one observation at a time and redoing the estimation. That is, let $\theta_{[-i]}$ denote the estimator with observation i deleted, for $i = 1, 2, ..., n$. Let

$$\bar{\theta}_{[-\cdot]} = n^{-1} \sum_{i=1}^{n} \theta_{[-i]} \qquad (3)$$

be the sample mean of these estimators. The Quenouille-Tukey Jackknife estimator (see, for example, ref *15*) is

$$\tilde{\theta} = n\hat{\theta} - (n-1)\bar{\theta}_{[-\cdot]} \qquad (4)$$

and can be shown to eliminate a bias term of order $1/n$. An estimator for the variance of the jackknife estimator is

$$\hat{\sigma}^2(\tilde{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} (\theta_{[-i]} - \bar{\theta}_{[-\cdot]})^2 \qquad (5)$$

Based on the assumed normal distribution for the estimator, a 95 confidence interval might be taken as

$$\tilde{\theta} \pm t\hat{\sigma}(\tilde{\theta}) \qquad (6)$$

where the multiplier 1.96 might be used if the assumption of normality holds, or we might use the value implied by the t-distribution, with degrees of freedom chosen on the basis of simulation. For example, twice the ratio of the squared mean of $\hat{\sigma}^2(\tilde{\theta})$ to its sample variance is sometimes used to estimate the equivalent degrees of freedom for the estimator. We found that this procedure tended to underestimate the degrees of freedom in simulations, leading to higher than nominal coverages for the confidence intervals.

The possibility that the observed data are contaminated by outliers or that the observations may have come from a mixture of distributions should be recognized. Every attempt should be made to isolate a specific cause for apparent outliers since the heavy-tailed distributions that dominate the water quality measurements tend to produce large values that may only look like outliers. If no irregularities due to sampling or analysis are obvious, we might evaluate the influence of the potential outlier by comparing the estimated parameter $\hat{\theta}$ to its value with the potential outlier deleted, say $\theta_{[-i]}$. This could be done by comparing the difference $|\theta - \theta_{[-i]}|$ to the expected standard deviation of the estimated difference, say $\hat{\sigma}(\hat{\theta})$. If the difference is greater than four or five standard deviations, the observation might be deleted.

Maximum Likelihood Estimation and the Box-Cox Transformation. Maximum likelihood estimation is also based on the assumption that the transformed variables are normally distributed, noting that the transformation (1) with the $y_i(\lambda)$, $i = 1, 2, ..., n$ assumed to be normally distributed will produce the log likelihood function

$$\ln L(\lambda, \mu_y, \sigma_y^{\,2}) \propto -\frac{n_1}{2} - \frac{1}{2\sigma_y^{\,2}} \sum_{y_i > U_i} (y_i - \mu_y)^2 + \sum_{y_i > U_i} (\lambda - 1) \ln x_i + \sum_{y_i \le U_i} \ln \Phi(Z_i) \qquad (7)$$

for the untransformed $x_i(\lambda)$, where the sums run over the uncensored, $y_i > U_i$, and censored, $y_i \le U_i$, observations, respectively, with the U_i defined as the transformed detection limits. Note that the argument of the normal cumulative distribution function is the transformed and standardized variable given by eq 16 of the Appendix. Box and Cox (*19*) proposed evaluating the log likelihood for each power λ and choosing the transformation for which the log likelihood is maximized. This requires estimating μ_y and σ_y^2 for each λ, which can be done recursively using the Expectation Maximization (EM) algorithm, as in ref *10*.

To apply the EM algorithm, let the current estimators be denoted by μ_y', $\sigma_y'^2$. Intuition suggests that the next estimators for the mean should depend on the sample mean of the observed data and the average of the expectations of the censored data, taken conditionally on being less than the detection threshold U_i for the ith censored observation. It turns out that the EM algorithm defines an iterative sequence involving re-estimating the mean and variance at each step. The mean is updated by averaging the observed observations with the conditional means of the censored observations evaluated at the previously estimated parameter values, i.e.

$$\hat{\mu}_{\mathrm{y}} = n^{-1}\Big(\sum_{y_i \le U_i} E(y_i|y_i \le U_i) + \sum_{y_i > U_i} y_i\Big) = n^{-1}\Big(\sum_{y_i \le U_i} (\mu' - \sigma_y' R_i') + \sum_{y_i > U_i} y_i\Big) \qquad (8)$$

Note that the corrected mean in eq 8 is just the original mean corrected by a scaled version of the ratio of the normal density to the cdf, as given in eqs 14−16 of the Appendix.

For the updated variance, it turns out that we simply average the usual squared residuals from the previously estimated mean and the conditional form of the previously estimated variance, i.e.

$$\hat{\sigma}_y^{\,2} = n^{-1}\Big(\sum_{y_i \le U_i} \mathrm{var}\,(y_i|y_i \le U_i) + \sum_{y_i > U_i} (y_i - \hat{\mu}_y)^2\Big) = n^{-1}\Big(\sum_{y_i \le U_i} \sigma_y'^2 (1 - Z_i' R_i') + \sum_{y_i > U_i} (y_i - \hat{\mu}_{\mathrm{y}})^2\Big) \qquad (9)$$

where R_i' and Z_i' are defined as eqs 15 and 16 evaluated at μ_y' and σ_y'. Note that each component of the first sum is the conditional variance as defined by eq 17. This simple updating procedure is repeated, with the new estimators $\hat{\mu}_y$ and $\hat{\sigma}_y$ taking on the role of μ_y' and σ_y', with the log likelihood eq 7 monitored for convergence. The EM algorithm is guaranteed to increase the log likelihood eq 7 at each step, and the log likelihood converges to a unique maximum when one exists. Furthermore, the algorithm is robust to starting values, which can be chosen as the sample mean and variance of the observed values or set at arbitrary values. The variances and covariances of $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ are given in ref *10* as the elements of the inverse of the negative of the information matrix, I. The information matrix has components that are relatively simple functions of the quantities Z_i, R_i and the true mean and variance. However, our simulations show that the jackknife procedure described in the previous section produces variances that lead to better coverages for the confidence intervals.

Of course, the estimated means, μ_x, and the estimated variances of the means in the untransformed scale are of primary interest. Shumway et al. (*10*) found that a limited set of Box-Cox values $\lambda = 0$, .50, 1.00 were effective for environmental data. Recall that 0 generates the log-normal distribution, and 1 assumes the normal distribution with the intermediate value .5 corresponding to the square root. Shumway et al. (*10*) have also investigated the success rate for Box-Cox in choosing the correct transformation. They found that the method works well when choosing no transformation (70% correct) and in choosing the log-normal transformation (48−68%) but has more difficulty in choosing the square root (50%) on data with known transformations. Hence, simply finding the transformation that maximizes the log likelihood for each particular data set and then applying that transformation is not necessarily recommended. Rather, it might be more appropriate to determine the transformation by a vote over a particular class of data sets and then apply the most popular transformation across all data in the class.

In the case of the logarithmic and square root transformations, the means for the untransformed data will be given by

$$\mu_{\mathrm{x}} = \begin{cases} \exp\left\{\mu_y + \frac{1}{2}\sigma_y^{2}\right\}, & \lambda = 0 \\ \left(\frac{1}{2}\mu_{\mathrm{y}} + 1\right)^2 + \frac{1}{4}\sigma_y^{2}, & \lambda = .50 \\ \mu_y + 1, & \lambda = 1.00 \end{cases} \tag{10}$$

The above expressions will give the exact maximum likelihood estimators for the mean $\hat{\mu}_x$ in the original scale if they are evaluated at $\hat{\mu}_y$, $\hat{\sigma}_y$, but they are still nonlinear functions of the parameters in the transformed scale. Shumway et al. (*10*) expand the nonlinear functions of the above in a first-order Taylor series and obtain the approximate variance via the delta method. The above reference also considered Efron's bootstrap, Efron (*23*) and the bias corrected bootstrap of Efron (*24*), and found that the delta method performed better in simulations.

Beginning with ref *11*, considerable effort has been expended in finding an appropriate bias correction for the censored case and for $\lambda = 0$, i.e., the log-normal assumption Schneider and Weissfield fitted the Saw correction by least squares (see also ref *14*) and obtained a bias correction of the form

$$\hat{\mu}_{\mathrm{x}} = \hat{\mu}'_x + \frac{\hat{\sigma}'_x}{n+1}\exp\left\{2.692 - 5.439\,\frac{n_1}{n+1}\right\} \tag{11}$$

We note in succeeding simulations that some situations lead to positively biased estimators for the mean for censoring levels in excess of 50%, so that the positive correction above would actually increase the bias. It is also the case that the Saw correction was proposed for the Type II censoring situation rather than for the Type I situation considered in this paper.

An Example of MLE and ROS Methods. A sample of $n = 25$ log-normally distributed observations was generated, yielding $n_0 = 7$ censored observations. The true mean and variance in the original scale were $\mu_x = 2.77$ and $\sigma_x^2 = .56$, respectively. In general, the proposed procedure is to first study the properties of the various power transformations to find the best candidate for the class of data that are under consideration. This involves evaluating the log likelihood eq 7 over a range of powers, defined by the basic power transformation eq 1. For each λ, $0 \leq \lambda \leq 1$, the EM algorithms 8 and 9 are applied repetitively to estimate μ_y and σ_y^2. A plot of the log likelihood is shown in Figure 1, and we note that the maximum occurs for $\lambda = 0$, implying the log-normal distribution.

To continue, we apply the EM algorithm to the data with $\lambda = 0$ and starting values, $\hat{\mu}'_y = 1.11$, $\hat{\sigma}'_y = 1.344$, determined as the mean and variance of the uncensored sample data. Note that this simply involves computing successively adjusted means and variances using eqs 8 and 9. After 20 iterations of the EM algorithm, we obtain the final exact MLE $\hat{\mu}_y = 1.0321$ and the inverse transformed MLE $\hat{\mu}_x = 2.85$ from eq 10. Figure 2 shows a plot of the log likelihood surface, with the maximum indicated, and we note that the surface is well behaved with a unique maximum.

Note that the mean is over-estimated (2.85 vs 2.77) and that a 95% confidence interval, using the delta method of Shumway et al. (*10*), gives (2.61, 3.02) and covers the true value of 2.77. We evaluate the effectiveness of such confidence intervals in subsequent simulations and compare to the jackknife method that is favored in this paper.

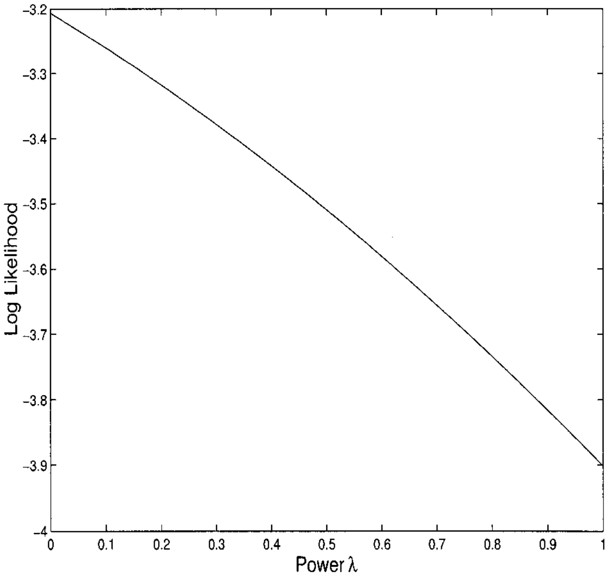

FIGURE 1. Log likelihood as a function of the Box-Cox power parameter λ.

Following through the first method of the section involves a regression on the order statistics (ROS) corresponding to the normal scores of the logarithms, followed by an extrapolation and then a computation of the conventional estimators of the adjusted sample, yielding ($\hat{\mu}_x = 2.82$, $\hat{\sigma}_x = .53$). The bias of these estimators is less than for the MLEs although the 95% confidence interval (2.65, 3.04) is very comparable.

Simulations

It is natural to ask what the limits might be on sample size and censoring level for data that are similar to that considered in the example of the previous section. Configurations of interest in this particular study were sample sizes on the order of $n = 20, 50$, with censoring of 50, 80% of the observed values. It is of interest to examine the bias, variance, and confidence coverage properties of the MLE and ROS procedures for these four experimental configurations. Also, the empirical distributions of the estimated means and the coverage properties of confidence intervals, computed from these estimators are of interest. Because the log-normal distribution is likely to be of greatest interest, we ran several simulations with $\lambda = 0$.

Table 1 shows the results of 500 repetitions of the process of simulating samples in the four basic experimental configurations. In this case, we considered maximum likelihood with the Box-Cox transformation (BC-MLE) and regression on order statistics both with (BC-ROS) and without (ROS) the transformation. Note that running the ROS method without a transformation is a rough method for evaluating its robustness to the log-normal departure. For the small samples, several observations can be made. First, the bias of the MLE is small and positive, about 5% of the mean, whereas the bias of the ROS is essentially zero. Note that the jackknife estimator for the MLE reduces the bias to essentially zero in all cases. The standard deviations of both ROS estimators are slightly smaller than the MLE, implying that the confidence intervals will be tighter at a given significance level. The coverage properties of all three estimators match the nominal coverage of 95%. For the larger sample, the three methods give comparable results; the bias of the maximum likelihood estimator is about 2%. Note that the excellent performance of the ROS estimator without any transforma-

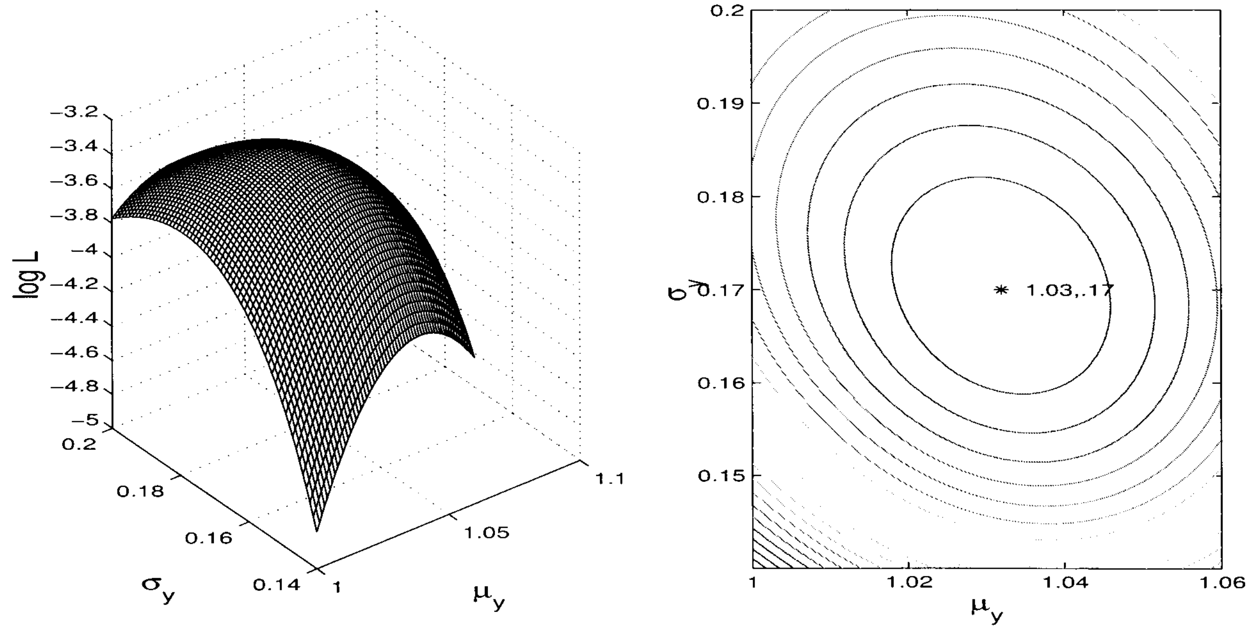

FIGURE 2. Log likelihood surface for contrived data.

TABLE 1. Average of Estimated Means, Standard Deviations (SD), and Jackknife Means, Mean-J (SD) for Log Normal Population with $\mu_x = 2.77$, $\sigma_x = 0.56$[a]

n (% censoring)	method	mean (SD)	mean-J (SD)	95% coverage
20(50)	BC-MLE	2.92(0.15)	2.76(0.14)	95
	BC-ROS	2.77(0.12)	2.77(0.12)	94
	ROS	2.77(0.12)	2.77(0.12)	93
20(80)	BC-MLE	2.93(0.15)	2.75(0.14)	96
	BC-ROS	2.77(0.12)	2.77(0.12)	95
	ROS	2.77(0.12)	2.77(0.12)	93
50(50)	BC-MLE	2.83(0.08)	2.78(0.08)	97
	BC−ROS	2.83(0.08)	2.78(0.08)	94
	ROS	2.77(0.08)	2.77(0.08)	94
50(80)	BC-MLE	2.81(0.08)	2.77(0.08)	96
	BC−ROS	2.77(0.08)	2.77(0.08)	94
	ROS	2.77(0.08)	2.77(0.08)	94

[a] Methods are exact maximum likelihood (MLE) and robust regression on normal scores (ROS) with Box-Cox (BC-MLE, BC-ROS) and without Box-Cox (ROS) for samples of size $n = 20$, 50, and 50%, 80% censoring, 500 replications.

tion modification at all in all cases implies robustness to the log-normal departure in both sample sizes. The biases predicted by the Saw correction (eq 13) were small and negative (0−5%) and would have adjusted in the wrong direction.

To test the consistency of the above results, other simulations were run which yielded similar values except for the magnitude of the biases of the unadjusted MLE. For example, increasing the theoretical mean and standard deviation to $\mu_x = 2.276$, $\sigma_x = 1.213$ again produced maximum likelihood estimators that were biased about 15% high in small samples with the jackknifed, MLE again having essentially zero bias. There were also biases in the large sample (7%) by the MLE, which went to zero with the jackknife correction. The ROS method was basically unbiased in all cases and had substantially smaller standard deviations, .26 compared to .34 in the smaller samples, and .17 compared to .18 in the larger samples. Again, the nominal coverages were attained for all estimators.

Empirical distributions of the MLE and ROS estimators, both with and without the jackknife, are shown in Figures 3 and 4 for the small sample size and 50% censoring. We note the generally symmetric nature of the empirical distributions in both cases that are approximately normal and are clustered around the theoretical value $\mu_x = 2.77$. The sampling distributions of the standard deviations are shown in Figures 3 and 4 for the maximum likelihood and normal scores estimator, with samples of size 20 and 50 censoring. It is clear that the delta method (left panel) and jackknife estimators (right panel) for the distributions of the MLEs are comparable, and we can use the asymptotic delta results. Similarly, the naive estimator for the sample variance s^2/n in the left panel of Figure 4 is not that much different from the jackknife estimator in the right panel, verifying that what we think might be used in the literature is reasonable. The average standard deviation estimators, shown in Table 1 for the log-normal case, are also comparable, whether we use the asymptotic or jackknife estimator for the MLE or the standard s^2/n or the jackknife estimator for the ROS estimator. We obtained comparable results for a small sample ($n = 20$) with high censoring (80%). The distribution of the MLE estimators look approximately normal, whereas the ROS estimators are somewhat skewed, with relatively long left tails. Comparable results showing the sampling distributions for the larger samples ($n = 50$) and the 80% censoring levels were similar.

In general, using a multiplier of $t = 2$ in eq 6 seemed to produce the best agreement between nominal and actual coverage for the 95 confidence intervals. As mentioned earlier, we experimented with estimating the equivalent degrees of freedom of the jackknife variance and found, in simulations, that the resulting estimators were too low and produced intervals that were too conservative.

The above results are relevant for the case where it is known that there exists a Box-Cox transformation, namely, the logarithmic transformation, that produces exactly normal data in the transformed domain. It is natural to wonder what would happen if there were no exact transformation to normality within the Box-Cox range. With this in mind, data were generated from a gamma distribution with mean value 4.00 and standard deviation 2.83. Searching the log likelihood eq 7 for the appropriate Box-Cox transformation when the data were generated from a gamma distribution tends to produce a maximum close to $\lambda = 5$ for the four cases, and this value was settled on as an approximation to a trans-

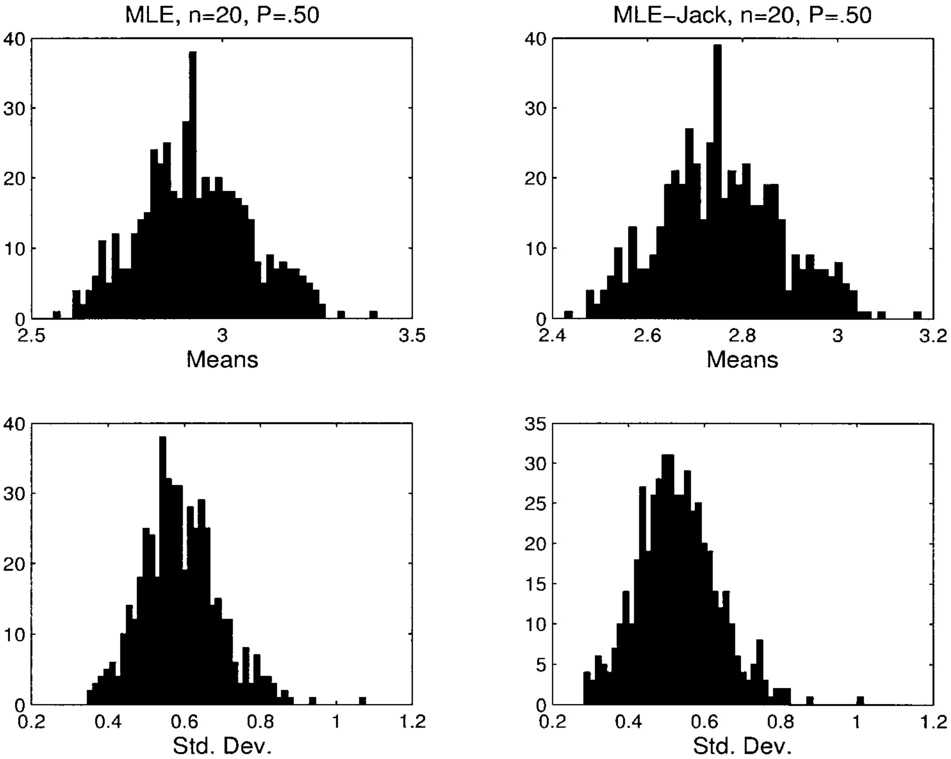

FIGURE 3. Empirical distributions of maximum likelihood and jackknife MLE estimators for means (upper panels) and standard deviations (lower panels) in samples of size 20 with 50% censoring, 500 repetitions.

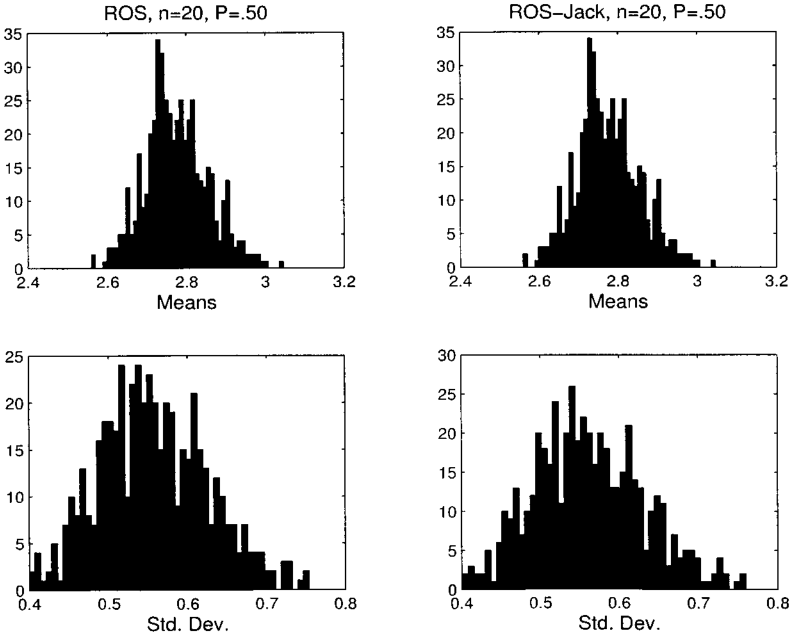

FIGURE 4. Empirical distributions of ROS and jackknife ROS estimators for for means (upper panels) and standard deviations (lower panels) in samples of size 20 with 50% censoring, 500 repetitions.

formation to normality. Table 2 presents comparable simulation results under this scenario.

Table 2 shows results that vary quite a bit from those obtained in Table 1. In this case, MLE and ROS with the Box-Cox transformation are comparable, with slight negative biases and confidence interval coverages that are somewhat less than the nominal 95%. In particular, the jackknife still seems to improve the performance of the maximum likelihood estimator but actually produces more bias in the ROS estimator. The ROS without the Box-Cox transformation is heavily biased, generally from 30 to 50%, over all configurations and produces an estimator with a large variance. Hence, the untransformed ROS estimator is distinctly nonrobust against the gamma distributed data and cannot be recommended for use in this case. In general, the MLE seems to be superior in this case because of smaller bias and better coverage.

Water Quality Data Analysis

Analyzing data that are likely from typical measurements is important, and we briefly consider some typical concentration data for the metals copper, aluminum, arsenic, chro-

TABLE 2. Average of Estimated Means, Standard Deviations (SD), and Jackknife Means, Mean-J (SD) for Gamma Distributed Populations with $\mu_x = 4.00$, $\sigma_x = 2.83$[a]

n (% censoring)	method	mean (SD)	mean-J (SD)	95% coverage
20(50)	BC-MLE	3.95(0.67)	3.94(0.66)	92
	BC-ROS	3.96(0.63)	3.81(0.75)	91
	ROS	2.90(0.86)	2.75(1.41)	91
20(80)	BC-MLE	4.48(0.92)	4.31(0.94)	89
	BC−ROS	4.57(0.62)	3.43(1.86)	85
	ROS	2.40(1.02)	2.03(4.38)	96
50(50)	BC-MLE	3.91(0.43)	3.90(0.43)	93
	BC-ROS	3.90(0.40)	3.87(0.45)	94
	ROS	2.80(0.58)	2.72(0.86)	72
50(80)	BC-MLE	4.07(0.63)	3.93(0.59)	95
	BC-ROS	3.98(0.40)	3.63(0.98)	89
	ROS	2.93(0.57)	2.86(0.82)	85

[a] Methods are exact maximum likelihood (MLE) and robust regression on normal scores (ROS) with (BC-MLE, BC-ROS) and without Box-Cox for samples of size n = 20, 50 and 50%, 80% censoring, 500 replications.

TABLE 3. Analysis of Metal Concentrations

method	l	n (% censored)	mean	SD	95% conf int
copper	0.10				
MLE		346(2)	15.98	14.53	14.48−17.47
ROS		346(2)	15.85	14.28	14.35−17.35
aluminum	0.00				
MLE		41(17)	118.26	218.28	59.17−177.36
ROS		41(17)	152.87	398.16	31.72−274.02
arsenic	0.00				
MLE		75(36)	2.13	2.41	1.61−2.66
ROS		75(36)	2.15	2.33	1.64−2.65
chromium	0.00				
MLE		336(43)	2.79	2.12	2.57−3.02
ROS		336(43)	4.74	13.17	3.41−6.05
nickel	0.10				
MLE		335(76)	4.16	5.22	3.60−4.73
ROS		335(76)	5.74	13.87	4.35−7.13
lead	0.00				
MLE		335(9)	5.50	12.40	4.41−6.59
ROS		335(9)	7.32	23.48	4.85−9.79

mium, nickel, and lead. The data are from the California Department of Transportation (Caltrans) Stormwater Management Program related to the highway runoff characterization monitoring (1997−1999) under the National Pollution Discharge Elimination System. Table 3 gives the particulars for sample sizes that ranged from $n = 75$ to $n = 346$ and censoring levels varying from 2 to 76%. To look at the distributions of concentrations, we plotted the log likelihood eq 7 as a function of the power λ in the transformation eq 1 and obtained again $\lambda = .10$ for copper and nickel and $\lambda = 0$ for all other sets, implying that the logarithmic transformation would be the best positive power. Figure 5 shows the plot for copper, and it is clear that the log likelihood rises montonically to a maximum at .1, which is close to zero.

Figure 6 shows the histograms for the log-transformed metal concentrations, and we note that the distributions appear to be roughly normal when we account for the data missing because of being below the detection limit. Including these data would give a huge component at the lower end of some of the distributions. The copper distribution is based on a relatively large sample (346) with a small amount of censoring (2%) and seems to be close to normally distributed. The distributions of aluminum, arsenic, and nickel are less convincing. We also note the possible presence of outliers in all of these data sets.

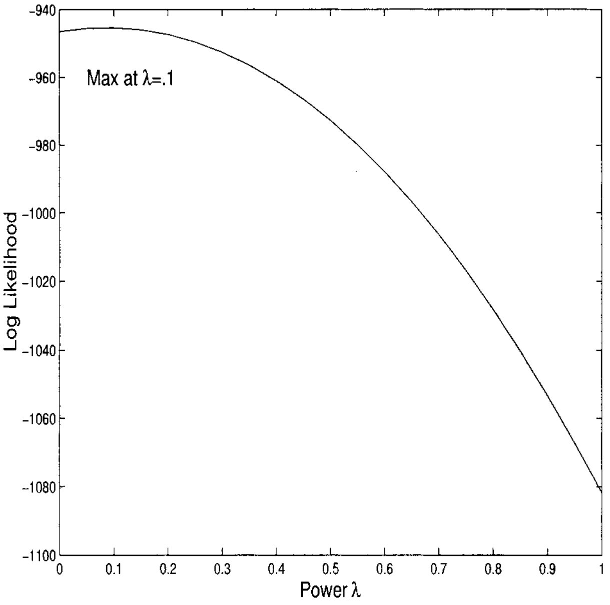

FIGURE 5. Log likelihood as a function of the Box-Cox power parameter λ for copper data.

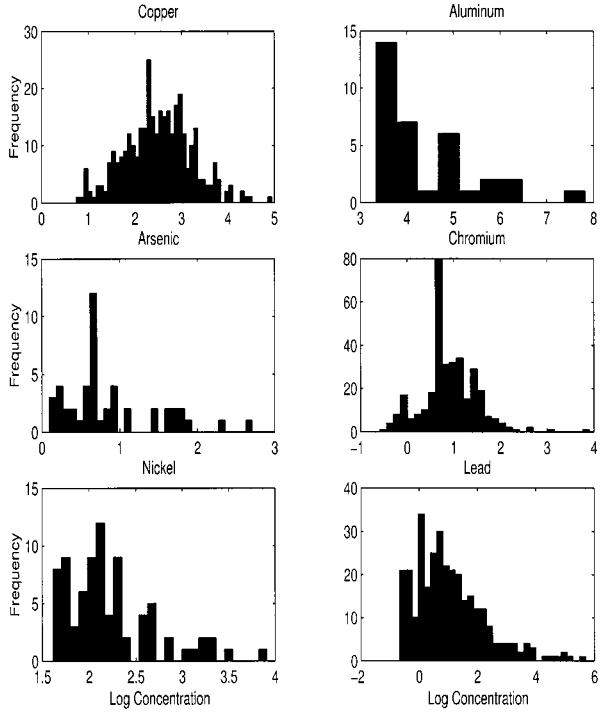

FIGURE 6. Histograms for copper, aluminum, arsenic, chromium, nickel, and lead data (see Table 3).

Table 3 shows the estimated mean concentrations for the six metals. We note that there is rough agreement of the MLE and ROS estimators except for the aluminum concentration data. The aluminum data has a possible outlier with a concentration of 2500 as compared to the next highest measured concentrations of 540 and 570. The MLE method yields 118.26(30.15) for the mean with the observation included and 89.54(18.38) with the observation excluded, whereas the ROS estimator gave comparable values of 152.87-(61.81) and 94.65(20.91). It is clear in this case that deleting the outlier produced more compatible estimators for the mean for the two procedures.

Discussion

This paper has investigated the estimation of mean and variance parameters for severely censored small samples of non-normal water quality data. The effectiveness of several procedures in the literature based on exact maximum likelihood estimation (MLE) for censored non-normal data and a nonparametric procedure that performs a linear regression on order statistics (ROS) were tested under both simulated conditions and with real data involving measured metal concentrations. The jackknife, also suggested by Singh et al. (*20*), is developed as a tool for estimating the variances and reducing the bias of both estimators.

We conclude that neither method is consistently better, as measured by bias and the overall coverage properties of the 95 nominal confidence intervals. For log-normal populations, our simulations indicate that the exact maximum likelihood methods are biased for small severely censored samples but that the bias can be eliminated by the jackknife. Bias corrections applied in the past water quality literature were actually found to increase the bias in simulation experiments. For log-normal data, the ROS estimators were essentially unbiased for all test configurations and had slightly smaller standard errors, even when no transformation was applied. For gamma distributed populations and a square root transformation, both methods were biased, with the maximum likelihood estimator performing slightly better. Applying the ROS estimator without transforming the data for the gamma distribution gave highly biased estimators and large variances; the ROS estimator is not robust against the distributional departure in this case.

In the analysis of real water quality concentrations, it is noted that the MLE and ROS methods will give similar results for moderately skewed data but may produce quite different results for highly skewed data. The 95% intervals obtained from the MLE were much smaller than those obtained from ROS when the distributions were highly skewed. This was probably due to the linear model being a poor fit for the transformed order statistics.

In closing, it should be noted that the recommended procedure that is implicit from the conclusions of the paper is rather difficult to put into practice in an operating environment. While the recommended methodologies, based on either maximum likelihood or regression on order statistics, are simple to apply, the choice will depend on the particular configurations of censoring, sample size, and non-normality encountered in a particular database. We recommend grouping data into similar subsets and then applying the same Box-Cox transformation to all members, restricted to square root or logarithm and based on searching the log-likelihood over all samples in the subgroup. The choice between MLE and ROS depends somewhat on the kind of data encountered. MLE will be biased and ROS will not be biased for log-normal data, but the MLE bias will be reduced by the jackknife ROS will be robust in the log-normal case but may be severely biased in other contexts. Shumway and Azari (*25*) give software written in MATLAB that can perform all computations mentioned in this report, and subsequent versions are being developed under contract that will be menu driven and easy to apply.

Acknowledgments

This study was funded by Caltrans Stormwater Management Program under the contract number 43A0014. The support of the program chief, Mr. Steven Borroum, is gratefully acknowledged. We would also like to acknowledge the contributions of three reviewers and Associate Editor, Professor Mitchell Small. In particular, their incisive comments led to a substantial revision that modified some of the conclusions presented in the original version of the paper.

Appendix

We summarize for completeness some of the details for the equations given in the text. We use repeatedly in the main text the expressions for the standard normal density

$$\phi(z) = \frac{\exp\{-z^2/2\}}{\sqrt{2\pi}} \tag{12}$$

and the normal cumulative distribution function

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)dz \tag{13}$$

along with its inverse $\Phi^{-1}(P)$,defined as the solution of the equation $P = \Phi(x)$.

For the censored case, it is easy to show that the conditional expectation of the random variable y, given that $y \leq U$, the lower detection threshold, is given by

$$E(y|\ y \leq U) = \mu_{\mathrm{y}} - \sigma R \tag{14}$$

where R is the ratio

$$R = \frac{\phi(Z)}{\Phi(Z)} \tag{15}$$

and

$$Z = \frac{U - \mu_y}{\sigma_y} \tag{16}$$

is the current standardized residual. The conditional variance becomes

$$var(y|\ y \leq U) = \sigma_{\mathrm{y}}^{2}(1 - ZR) \tag{17}$$

The conditional means and variances are important components of the simple algorithm described in the text for computing the maximum likelihood estimators, i.e., the estimators that maximize the likelihood function given in the text.

The covariance matrix of the maximum likelihood estimators is given in ref *10* as the inverse of the negative of the information matrix, I, say $cov = (-I)^{-1}$. The elements of the 2×2 matrix I are the second partial derivatives with respect to μ_y and σ_{y}^2 of the log likelihood function eq 7. The entries in I are functions only of Z_i, R_i, μ_y and σ_y^2 and are given in ref *10*. To find an approximate large-sample variance of the means, μ_x in the original scale the above authors use the delta method and eq 10.

Literature Cited

(1) Hald, A. *Skandinavisk Aktuarietidskrift* **1949**, *32*, 119.
(2) Aitchison, J.; Brown, J. A. C. *The Log-normal Distribution*; Cambridge University Press: New York, 1957.
(3) Ringdal, F. *Bull. Seismolog. Soc. Am.* **1976**, *66*, 789.
(4) Cohen, A. J. *Technometrics* **1959**, *1*, 213.
(5) Persson, T.; Rootzen, H. *Biometrika* **1977**, *64*, 123.
(6) Newman, M. C.; Dixon, P. M.; Looney, D. B.; Pinder, J. E. III *Water Res. Bull.* **1989**, *25*, 904.
(7) Ahn, H. *J. Am. Water Resources Assoc.* **1996**, *34*, 583.
(8) Newman, M. C.; Green, K. D.; Dixon, P. M. *UNCENSOR, Version 40*; Savannah River Ecology Laboratory: Aikin, SC, 1995.
(9) Shumway, R. H; Azari, A. S. *Estimating mean concentrations when some data are below the detection limit*; Final Report, A733-045; California Air Resources Board, 1988.
(10) Shumway, R. H.; Azari, A. S.; Johnson, P. *Technometrics* **1989**, *31*, 347.
(11) Saw, J. G. *Biometrika* **1961**, *64*, 123.
(12) Schneider, H.; Weissfield, L. *Biometrics* **1986**, *42*, 531.
(13) El-Shaarawi, A. H. *Water Resources Res.* **1989**, *25*, 685.
(14) Haas, C. N.; Scheff, P. A. *Environ. Sci. Technol.* **1990**, *24*, 912.

(15) Efron, B. *The Jacknife, the Bootstrap and Other Resampling Plans*; CBMS/NSF Monograph 38, Society for Industrial and Applied Mathematics: Philadelphia, 1989.
(16) Gilliom, R.; Helsel, D. *Water Resources Res.* **1986**, *22*, 135.
(17) Helsel, D. R.; Gilliom, R. J. *Water Resources Res.* **1986**, *22*, 147.
(18) Gupta, A. K. *Biometrika* **1952**, *39*, 260.
(19) Box, G. E. P.; Cox, D. R. *J. Royal Statist. Soc., B* **1964**, *39*, 211.
(20) Singh, A. K.; Singh, A.; Engelhardt, M. EPA Technology Support Center Issue, 1997; December, 1−19.
(21) Gleit, A. *Environ. Sci. Technol.* **1985**, *19*, 1201.
(22) Blom, G. *Statistical Estimates and Transformed Beta Variables*; Wiley: New York, 1958; pp 68−75, 143−146.
(23) Efron, B. *Annals Statistics* **1979**, *7*, 1.
(24) Efron, B. *Biometrika* **1985**, *72*, 45.
(25) Shumwayl, R. H.; Azari, A. S. *Statistical approaches to estimating mean water quality concentrations with detection limits*; Final Report, Contract 43A0014; Task 33, California Department of Transportation, Department of Statistics, University of California, Davis, 95616, 2000.

Received for review July 5, 2001. Revised manuscript received May 8, 2002. Accepted May 10, 2002.

ES0111129